A Corpus-based Study of Lexical and Grammatical Features of Written Business English

An M.A. thesis
submitted to the Graduate Department of Language and Information Sciences
of
the University of Tokyo
in partial fulfillment of the requirements for the degree of master of art
by

Yasumasa Someya

Chief Supervisor: Professor Hideo Suzuki
Second Supervisors: Professors Hideo Oka and Kazuhiko Matsuno

1999

SUMMARY

This study was conducted with the primary purpose of identifying and describing some of the major lexico-grammatical features of English for Business Purposes (EBP) in its written domain. The author also aimed at establishing a computer-based study methodology appropriate for the analysis of such large-scale corpus data as those used in this study, including the development of a series of computer programs with which to analyze the corpus data from various viewpoints and for specific purposes.

In Chapter 1, the argument was made that relatively little serious research existed in the area of EBP and, as a result, we actually know surprisingly little about language usage and performance in international business contexts. One of the consequences of this general paucity of attested knowledge in the area was that most of the EBP teaching/learning materials currently available on the market were written largely on the basis of intuitive judgments of respective authors as to what make EBP unique from other

areas of ESP, or from EGP for that matter, in its lexical, grammatical and pragmatic aspects.

The present study was conceived against this background, with the study objectives as stated in the opening paragraph of this summary. In carrying out the study, the author established three working hypotheses as to the nature of EBP; i.e. the "lexical closure" hypothesis, the "lexical ease" hypothesis, and the "write-as-you-talk" hypothesis. It was hoped that, by setting these working hypotheses, this paper would have clear foci of discussion rather than being all too inclusive.

In the latter part of this chapter, the author defined the notion of "Business English" (BE). In his definition, BE refers to the kind of English "used for business purposes in international and inter-cultural contexts as a common means of communication by the people who do not share a first language" (p. 5). It was thus viewed primarily as a branch of ESP. This was followed by a review of some of the major research related to the lexical and grammatical features of EBP.

As a corpus-based, data-driven endeavor, the present study drew heavily on corpus data as a basis -- in fact, the sole basis -- for arguments. Also, in order to properly handle and analyze linguistic corpora as large as those used in the present study, the use of some computer programs and data-analysis software were considered indispensable. In the last section of this chapter, therefore, a rather detailed discussion as to the contents of the corpora and computer programs and software were presented in as much non-technical manner as possible. The corpora used in the present study were: (1) the Business Letter Corpus (BLC), which was the main study corpus; (2) three reference corpora consisting of the Brown, LOB and TIME Corpora, and (3) the Learner Corpus of English Business Letters (Learner BLC), which consisted of an adjusted total of about 205,000 words of business letters written by Japanese business people. The size of all the other corpora was about one million words (running tokens).

In Chapter 2, we turned to the analysis of the corpus data from a viewpoint of parts-of-speech (POS) distribution. After reviewing the main rationale for the analysis, Section 2.2 presented a brief description of the study procedure of the POS analysis. The analysis revealed, among other things, that nouns constituted the most prominent word class category, comprising an average of about 27.57% of all the word tokens in the study corpora. This was followed by prepositions (13.35%), verbs (11.18%), determiners (10.17%), adjectives (7.7%), adverbs (5.21%), and pronouns (7.21%). When the BLC and the three reference corpora were compared binominally for each of these POS categories, the most significant differences (e.g. significant if $Z > 3.29$, $p > 0.001$) were found in modals ($Z = 9.2513$), pronouns ($Z=6.0369$), infinitival to ($Z=5.8402$), determiners ($Z=5.8283$), conditional *if* ($Z=5.5106$), and wh-particles ($Z=5.2919$). The analysis thus indicated that these six

categories were particularly prominent in quantitative terms in the BLC, suggesting the need for more detailed probes into those categories.

We also compared the data between the Native and Learner BLCs, and found that the overall patterns of POS distribution of two corpora were very much the same, indicating that native and non-native English were after all not that different as to the proportions of nouns, verbs, prepositions and other parts of speech people use in their writing. Despite this overall similarity, some noticeable, if not statistically significant, differences were found in such areas as nouns, pronouns, determiners, adjectives and, to a lesser degree, in verbs and modals. It was argued that the relatively higher proportion of nouns observed in the Learner BLC was largely due to the effect of "nominalization," which, we presumed, would also explain both the lower proportion of verbs in the Learner BLC on the one hand, and the higher proportions of determiners and prepositions on the other. A more detailed discussion on this topic, however, was left for Chapter 4.

In Chapter 3, a total of six different types of wordlist compiled from the main study corpus were introduced and their contents briefly explained. First, in Section 3.1, some of the major technical terms and statistical concepts used in the wordlists, or needed to understand the contents thereof, were defined and explained in some detail. The main statistical concepts we employed to measure relative importance of each lexical item in the study corpus included those of *Usage* and *Keyness.* The former is a product of the actual frequency of a given lexical item in a corpus multiplied by the coefficient of dispersion of that item among the smaller subsets of that corpus. As such, it usually gives a better estimate of the relative importance of each lexical item in the entire corpus than simply resorting to frequency data alone. The *keyness* of word *x*, on the other hand, is calculated by comparing the frequency of that word in the target corpus with that of the same word in a much larger reference corpus, taking also into consideration the total numbers of running words (tokens) in both corpora. Ted Dunning's Log Likelihood test was used to calculate the statistical significance of a given *K-score*. This test "gives a better estimate of keyness (than does the classic Chi-squared test), especially when contrasting long texts or a whole genre against your reference corpus." (Scott, 1998. p. 65). In the current study, a word is considered to be "key" either positively or negatively if the *p* value obtained for that word is larger or equal to 0.000001.

Finally, in Sections 3.2 and 3.3, the six wordlists were presented and their technical contents explained as briefly as practically possible. The wordlists introduced in this part of the paper were: the COMPREHENSIVE BLC WORDLIST, the BLC KEYWORDS LIST, and four categorical wordlists covering verbs, adverbs, adjectives and nouns respectively. For the sake of simplicity, only the first page of each wordlist was included in this part of the paper

(i.e. Volume 1), except for the BLC KEYWORDS LIST, for which both the first and last pages were included. The full lists can be found in the second volume of the paper. In introducing the wordlists, we also reviewed the technical procedures of wordlist compilation to show not only the data we obtained, but also how they had been obtained.

In Chapter 4, we looked at the wordlists more closely from various viewpoints. First, in Section 4.1, the three working hypotheses set out at the outset of the paper were examined for their validity against the evidence presented in these wordlists. As to the first hypothesis, it was found that the type-token ratio of the BLC at the one-million-word ceiling was about 1:44, while that of the three reference corpora was about 1:22 on average, indicating that the BLC was about twice as lexically closed as the reference corpora. It was also found that we only need approximately 1,400 to 2,900 word-types to say about 90 to 95% of what we need to say in fairly formal written business discourse.

A separate analysis conducted on the Learner BLC further indicated that even these figures could be much too conservative from the viewpoint of non-native writers of business messages. The second hypothesis was also confirmed valid. In a nutshell, it was found that approximately 77.5% of the first 3,000 word-types in the BLC, which cover about 95.25% of all the BLC word tokens, were within the 4000 Basic Words as defined by the JACET (Japan Association of College Teachers), plus a fairly standard set of proper nouns, abbreviations and acronyms. The third hypothesis was tested against the corpus evidence with regard to those items referred to as significant in the foregoing POS analysis described in Chapter 2, including personal pronouns, articles, prepositions (*of*, in particular), and wh-particles (*which*, in particular). We also investigated the use of "contractions" (i.e. reduced forms of *will/shall, would/should, are,* and *am)* as it relates to the hypothesis in question. In all these instances, the evidence was in strong support of our assumption that written business English is characterized by its incorporation of spoken features into written texts.

In Section 4.2, we focused our attention on modals, which we found constituted the single most important lexical category in the entire BLC lexicon. First, we looked at the data more closely to find out exact information as to the within-the-category distribution of major modals. It was found that *will* and *would* constituted the two most important modals comprising about 60% of all the modal tokens in the BLC, with a Keyness score of 9,200 and 1,468 respectively. These two items were followed by, in Keyness order, *can* (K = 1036), *shall* (K = 385), *should* (K = 206), and *may* (K = 180).

We also discussed the syntactic environments in which modals appear, with particular reference to the data presented in Kennedy (1998), and found that the three syntactic frames, i.e. "MD + to VB", "MD + be + VBN" and "MD + be + (NP | AP | PP | .)," comprised

the canonical forms of modal verb-phrase and that the majority of modal tokens in the BLC occurred in these three structures. With regard to *if*-conditionals, which are closely related to modals, we came up with many interesting findings as summarized at the end of Subsection 3 of Section 4.2. Some of which are as follows: (1) About 80% of all the conditional sentences are those dealing with "open" conditions; (2) Patterns A (i.e. "If + VB, VB") and B (i.e. "If + VB, MDpres + VB") are the two main syntactic patterns of the "open" conditional, constituting about 76% of all the cases thereof; (3) Pattern D (i.e. "If + VBD, MDpast + VB") is by far the most frequent syntactic pattern of the "closed" conditional, constituting about 78% of all the cases thereof; and (4) Both Patterns E (i.e. "If + had VBN, MDpast + have VBN") and F (i.e. "If + were, MDpast + VB") are also used to indicate purely hypothetical conditions, but they are used only minimally.

In Section 4.3, we discussed the BLC vocabulary in more detail for the purpose of establishing the "core" vocabulary of Business English in each of the four word classes, i.e. verbs, adverbs, adjectives and nouns. In choosing the candidates for the core vocabulary, we adopted the two statistical measures, *Usage* and *Keyness*, described in Chapter 3. As to verbs (VB), we selected 350 entries covering about 90% of all the VB tokens in the BLC. For adverbs (RB), 300 entries covering 97% of all the RB tokens were selected. From these "core" items, we further identified 100 each of "key" keywords in the respective POS categories, and proposed that they be further investigated from both semantic and textual/discoursal perspectives for a more complete description of EBP lexicon. Concerning the remaining two categories, adjectives (JJ) and nouns (NN), a total of 300 entries for the former and that of 325 for the latter were tentatively selected as the primary candidates for inclusion in the core list.

From a pedagogical perspective, however, it was considered not quite enough, nor was it desirable, to simply provide the learners of EBP with a long list of "important words" for them to memorize. All the evidence show that a large part of their problems is not due to their lack of vocabulary as such, but stems mostly from insufficient understanding on the part of the learners as to the syntactic and semantic properties of individual lexical items that ultimately determine their idiosyncratic behavior. It was proposed, therefore, that the core list, for it to be truly useful, needs to be supplemented with the kind of information mentioned above, in the form of "lexical profiling" of which this author provided an example in Appendix C3.